# 零样本图像识别

兰 红* 方治屿

(江西理工大学信息工程学院 赣州 34100)

**摘 要**：深度学习在人工智能领域已经取得了非常优秀的成就，在有监督识别任务中，使用深度学习算法训练海量的带标签数据，可以达到前所未有的识别精确度。但是，由于对海量数据的标注工作成本昂贵，对罕见类别获取海量数据难度较大，所以如何识别在训练过程中少见或从未见过的未知类仍然是一个严峻的问题。针对这个问题，该文回顾近年来的零样本图像识别技术研究，从研究背景、模型分析、数据集介绍、实验分析等方面全面阐释零样本图像识别技术。此外，该文还分析了当前研究存在的技术难题，并针对主流问题提出一些解决方案以及对未来研究的展望，为零样本学习的初学者或研究者提供一些参考。

**关键词**：零样本学习；深度卷积神经网络；视觉语义嵌入；泛化零样本学习

**中图分类号**：TN911.73; TP391.41 **文献标识码**：A **文章编号**：1009-5896(2020)05-1188-13

**DOI**: 10.11999/JEIT190485

# Recent Advances in Zero-Shot Learning

LAN Hong FANG Zhiyu

(*School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou* 341000, *China*)

**Abstract**: Deep learning has shown excellent performance in the field of artificial intelligence. In the supervised identification task, deep learning algorithms can achieve unprecedented recognition accuracy by training massive tagged data. However, owing to the high cost of labeling massive data and the difficulty of obtaining massive data of rare categories, it is still a serious problem how to identify unknown class that is rarely or never seen during training. In view of this problem, the researches of Zero-Shot Learning (ZSL) in recent years is reviewed and illustrated from the aspects of research background, model analysis, data set introduction and performance analysis in this article. Some solutions of mainstream problem and prospects of future research are provided. Meanwhile, the current technical problems of ZSL is analyzed, which can offer some references to beginners and researchers of ZSL.

**Key words**: Zero-Shot Learning (ZSL); Deep Convolutional Neural Networks (DCNN); Visual-semantic embedding; generalized Zero-Shot Learning (gZSL)

## 1 引言

随着近年来深度学习在人工智能领域中的广泛应用，基于卷积神经网络的有监督学习在人脸识别、图像分类、图像重建等方面取得了优异成果。特别是在分类识别任务中，通过对海量带标签数据的训练，可以达到前所未有的识别精度(如人脸识别领域，DeepID2[1]在LFW数据集上精度达到了99.15%；图像分类领域，PNAS[2]在ImageNet数据上的精度分别达到top1=82.9%, top5=96.2%；图像重建领域，SRDenseNet[3]在图像超分辨率应用中取得PSNR=32.05 dB, SSIM=0.9019的成果)。然而对于海量标签数据的获取目前仍存在问题：(1)对海量数据进行标注工作成本昂贵，对每一个类别均要收集大量的训练数据，并且对这些数据逐一的进行人工标注既费时又费力；(2)对于某些罕见对象获取数据较难，例如获取珊瑚裸尾鼠(2019年2月灭绝)的图像数据；(3)在数据缺乏的情况下进行细粒度的物体识别更为困难。

但是相比之下，人类可以区分至少30000个基础物体类和更多的子类[4]，而且在认知科学领域中，有研究表明人类具有知识迁移能力，可以在大脑中保存、积累过去获得的知识，而且在学习新事物的时候能够提取目标物体的中间语义表示或更高级的

收稿日期：2019-07-01；改回日期：2019-11-03；网络出版：2019-11-13
*通信作者： 兰红 lanhong69@163.com
基金项目：国家自然科学基金(61762046)，江西省自然科学基金(20161BAB212048)

Foundation Items: The National Natural Science Foundation of China (61762046), The Natural Science Foundation of Jiangxi Province (20161BAB212048)

抽象描述，并将先前学习的知识迁移到目标物体。这也解释了为什么人类可以理解从未见过的全新事物或是只见过一两次就能准确再次识别该类事物。

可是对于传统机器学习和现在大多数有监督的深度学习来说却无法做到这一点，为了解决无带标签训练数据(未知类)的学习任务，许多学者受人类识别能力的启发，开始研究更加符合人类终身学习特点的系统，实现最终类人的人工智能。零样本学习是实现这一目标的关键一环，通过零样本学习可以解决目前对未知类的识别，让人工智能系统逐渐摆脱大量带标签训练数据的限制。

## 2　零样本学习介绍

零样本学习研究可以追溯到2008年，Larochelle等人[5]针对字符分类问题提出了零样本学习，并在当时将其命名为“zero-data learning”，主要研究当训练类与测试类互斥时的目标分类问题。2009年，Palatucci等人[6]将其扩展至神经信息处理系统领域，并更名为“zero-shot learning”。同年，Lampert等人[7]将零样本学习扩展至计算机视觉领域，提出经典的基于属性学习的算法，贡献出了目前在零样本学习性能评价中广泛使用的AwA (Animals with Attributes)数据集。至此，才算真正打开零样本学习的关注度。

### 2.1　基本概念

在没有任何观测数据的情况下进行分类识别任务被称为零样本学习，其关键思想在于探索和利用未知类与已知类在语义或其他高层特征间的相关知识，从而达到知识迁移的目的。零样本学习的形式化数学定义为：给定符号$\mathcal{X}$表示图像的特征空间，符号$\mathcal{Y}$表示已知类集合，符号$\mathcal{Z}$表示未知类集合，其中$\mathcal{Y}$和$\mathcal{Z}$之间互斥，即$\mathcal{Y}\cap\mathcal{Z}=\varnothing$，符号$\mathcal{D}_r$表示由带标签图像组成的训练数据集，有$\mathcal{D}_r=\{x_i,z_j\}_{i=1}^{N}$，符号$\mathcal{D}_\infty$表示由在训练数据集中从未出现且不带标签的图像组成的测试数据集，有$\mathcal{D}_\infty=\{x_j,z_j\}_{j=1}^{M}$，零样本学习的任务是使用$\mathcal{D}_r$来构建分类器$C$实现对$\mathcal{D}_\infty$的精确分类，即$C:\mathcal{X}\to\mathcal{Z}$。零样本学习最大的特点就是在模型测试阶段使用的数据集从未出现在模型训练阶段的数据集中，表1列举了在计算机视觉领域几类常见机器学习方法与零样本学习之间的差异比较。

无监督学习[8]和有监督学习[8]作为机器学习领域的两大核心，前者是通过大量的无标签数据训练分类器，利用样本数据间的内在关系实现分类，后者则是通过大量的带标签数据训练分类器，寻找特征与标签之间的联系，完成分类任务；半监督学习[9,10]是极端情况下的一种结合无监督学习与有监督学习的方法，因为大量带标签数据获取较难，所以学者通过研究发现使用少量带标签数据和大量无标签数据训练的分类器对已知类的识别性能有大幅提升，从而提出半监督学习；少样本学习[11]是比半监督学习更极端的情况，即训练数据中测试类只有一个或是极少数的带标签数据，在这种情况下完成分类任务的学习方法；零样本学习[12,13]则是比少样本学习更极端的情况，即在训练数据中完全没有测试类的带标签数据，甚至是完全没有测试类的任何数据，在这种情况下完成分类任务的学习方法。

### 2.2　主要类别

零样本学习不同于以往的有监督机器学习方法，因为用于测试的未知类与用于训练的已知类具有互斥关系，使用传统的有监督学习理念从带标签数据集中训练出来的分类器无法正确地完成对未知类的预测。为了克服该问题，零样本学习通过建立已知类与未知类的耦合关系，实现将已知类学习到的知识迁移到未知类上，从而达到对未知类的预测效果，图1给出了零样本学习的技术结构图。

从图1可以看出，零样本学习是通过嵌入空间来建立已知类与未知类之间耦合关系，这也是零样本学习的核心。它利用原始图像提取特征向量构建特征空间，再利用类别标签提取图像对应类别的语义向量构建语义空间，然后结合特征空间和语义空间，分析两者之间的映射关系，构建嵌入空间。基于语义空间，在训练阶段使用已知类数据集学习图像与类别之间的关系，在测试阶段就可以利用该关系，先由图像特征预测对应的语义向量，再根据语义向量匹配图像所属类别。零样本学习对应的形式化公式表达为

表 1　机器学习方法对比表

| | 训练集$\{\mathcal{X},\mathcal{Y}\}$ | 测试集$\{\mathcal{X},\mathcal{Z}\}$ | 训练类$\mathcal{Y}$与测试类$\mathcal{Z}$间关系$R$ | 最终分类器$C$ |
|---|---|---|---|---|
| 无监督学习 | 大量无标签图片 | 已知类图片 | $\mathcal{Y}=\mathcal{Z}$ | $C:\mathcal{X}\to\mathcal{Y}$ |
| 有监督学习 | 大量带标签图片 | 已知类图片 | $\mathcal{Y}=\mathcal{Z}$ | $C:\mathcal{X}\to\mathcal{Y}$ |
| 半监督学习 | 较少带标签图片和大量无标签图片 | 已知类图片 | $\mathcal{Y}=\mathcal{Z}$ | $C:\mathcal{X}\to\mathcal{Y}$ |
| 少样本学习 | 极少带标签图片和大量无标签图片 | 已知类图片 | $\mathcal{Y}=\mathcal{Z}$ | $C:\mathcal{X}\to\mathcal{Y}$ |
| 零样本学习 | 大量带标签图片 | 未知类图片 | $\mathcal{Y}\cap\mathcal{Z}=\varnothing$ | $C:\mathcal{X}\to\mathcal{Z}$ |

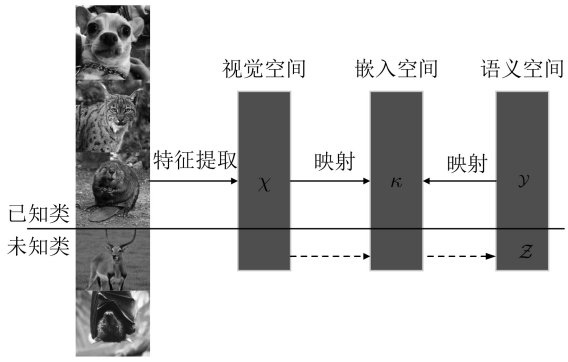



图 1 零样本学习技术结构图

$$\left.\begin{array}{l} C=l(f(x)) \\ f:\mathcal{X}\to\mathcal{K} \\ l:\mathcal{K}\to\mathcal{Z} \end{array}\right\} \tag{1}$$

其中，函数$f$是由训练数据集学习而来，用来完成图像特征$\mathcal{X}$到嵌入空间$\mathcal{K}$的映射，函数$l$一般选用最近邻算法或是其他更复杂的相似度计算方法，用来确定图像所属的类别标签$\mathcal{Z}$。目前根据在训练阶段是否使用未知类的无标记数据，可以将零样本学习分为两大类，分别是直推式零样本学习和归纳式零样本学习。

#### 2.2.1 直推式零样本学习

直推式零样本学习是指在分类器训练阶段用于测试的未知类中无标签数据是可用的，在训练阶段使用无标签未知类数据提高最终分类器对未知类的判别能力，如图2(a)所示。一方面，利用未知类中的数据训练分类器可以提高$f$在未知类上的泛化能力；另一方面，未知类与已知类的数据相互结合用于学习，可以联合优化分类器$C$。

直推式零样本学习最早是在2012年由Fu等人[15]提出，他们将一个生成模型引入用户定义的属性中，通过平均原型的K近邻来预测测试数据的属性分布。

#### 2.2.2 归纳式零样本学习

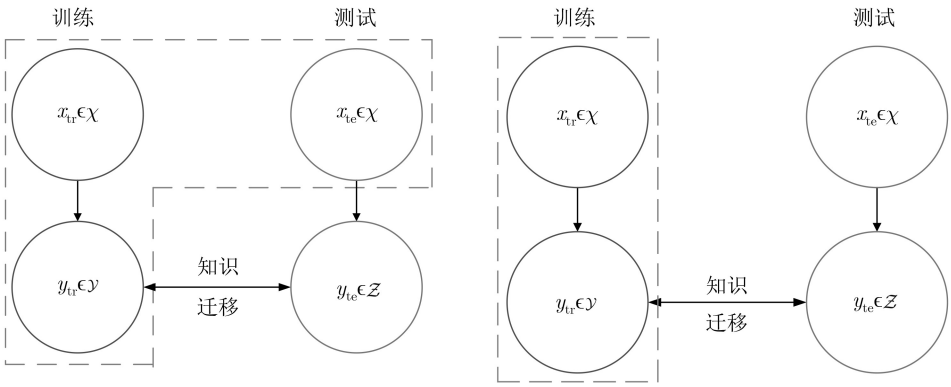
归纳式零样本学习是指在分类器训练阶段只使用训练集数据学习函数$f$，在测试阶段以并行方式对所有测试集数据进行类标签预测，预测过程相互独立，如图2(b)所示。考虑到直推式零样本学习的模型训练阶段对测试类的无标签样本具有硬性要求，一定程度上限制了零样本学习的模型训练过程，而归纳式的学习方式具有相对较好的灵活性和可扩展性，更符合实际问题的需求，所以该方法也是目前零样本学习问题中应用最广泛的一种解决思路。

在归纳式零样本学习中最经典的方法就是Lampert等人[7]提出的直接预测模型(Direct Attribute Prediction, DAP)和间接预测模型(Indirect Attribute Prediction, IAP)，它们分别采用两种不同的方式构建图片特征空间与语义嵌入空间的关系函数$f$。DAP模型使用训练数据直接学习图片特征到公共属性的映射，再由公共属性去匹配图片所属类别，如图3(a)所示；IAP模型则是使用已知类数据学习图片特征到已知类的映射，再构建类别到公共属性的映射，最后再由公共属性匹配未知类，如图3(b)所示。Lampert在2009发表的文章不仅为零样本学习问题提供了两种解决模型，更是启发了沿用至今的基于语义嵌入空间的零样本学习方法。

### 2.3 工程应用

从2009年掀起零样本研究热潮至今，已整整经历了10年时间，该项技术已经在众多领域中得到发展和应用。

在图像处理领域，Antol等人[16]通过抽象特定动作学习人与人之间的动作关系，以此来训练得到人类姿态识别模型，该模型实现抽象到现实的显性映射，可以精确地识别图片中人类的姿态；Robyns等人[17]受零样本学习的启发提出一种不依赖于局部和低维特征的指纹识别方法，实现在低廉的商用设备上就能达到高精度的识别效果；Palatucci等人[6]提出基于零样本学习的语义输出编码分类器(Semantic Output Code classifier, SOC)对fMRI图像进行解




(a) 直推式零样本学习示意图[14]　(b) 归纳式零样本学习示意图[14]

图 2 零样本学习示意图

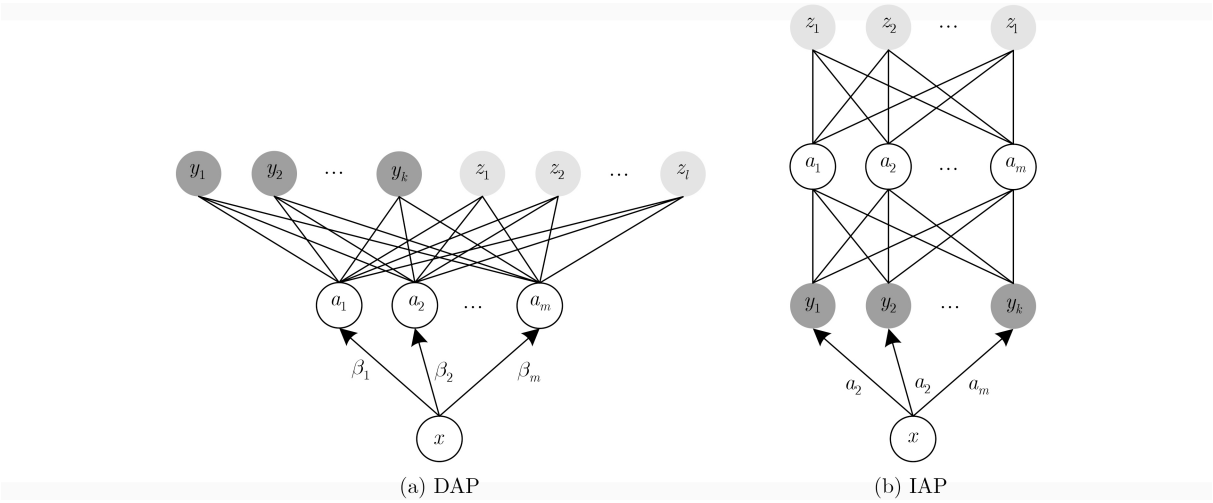



图 3 经典归纳式零样本模型示意图[7]

码分析，做到了对从未见过的图像的预测功能；Yang等人[18]提出Zero-shot Hashing算法将数据标签做二进制编码并投影到语义嵌入空间，利用已知类学习哈希函数，再将知识迁移到未知类，实现较以往哈希算法更优越的图像检索性能，同时规避了高昂的人工标签成本；文献[19]同样是为解决图像哈希算法中存在的真值注释缺乏问题，利用零样本学习将视觉空间中的图像特征和语义空间中的语义特征投影到一个通用的Hamming子空间中，并使用域自适应学习一个高质量的哈希函数。

在视频处理领域，Liu等人[20]使用基于属性的零样本学习将属性特征运用在人类行为识别上；Fu等人[15,21]为解决稀疏标签群体社会活动理解中的属性学习问题，利用零样本学习思想提出了一种学习潜在属性的模型；Jain等人[22]提出objects2action模型，利用零样本学习实现在视频中动作的时空定位；Xu等人[23]提出的辅助图像传输编码(Image Transfer Encoding, ITE)可以实现在辅助图像和文本语料库之间的知识迁移，解决了基于深度特征编码的大规模视频情绪识别中对未知情绪的识别问题。

在自然语言处理领域，Johnson等人[24]借助零样本学习在单神经网络下实现了对多语种的神经机器翻译(Neural Machine Translation, NMT)模型，提升了在训练源数据不充足时的翻译质量，单模型的设计不仅显著简化了以往的多语种NMT，还使得模型易于扩展到新的源-目的语言组合上；文献[25]使用一种简单的零样本学习方法解决了预测一个文档的多个以前从未出现过的可能标签的问题，实现多标签文本的分类任务。

零样本学习在各个领域都蕴藏着广阔的发展前景和巨大的应用潜力，随着学者们的深入研究，它将会让人工智能越来越类人化。

## 3　关键技术

目前学者们在处理零样本学习问题中使用的普遍的步骤为：(1)提取视觉特征构建视觉空间；(2)提取语义特征构建语义空间；(3)实现视觉空间与语义空间之间的映射构建嵌入空间。

### 3.1　视觉特征

在计算机视觉领域中，视觉特征提取是至关重要的一步，如何有效提取图片中最具价值，最能表示图像所描述内容的特征是计算机视觉长久以来的研究热点和难点。视觉特征一般包括颜色、纹理、形状等信息，传统的特征提取方法有描述颜色信息的HSV, Lab和HIS等颜色模型和HOG特征[26]，描述纹理信息的灰度共生矩阵, Tamura特征, 自回归模型和LBP算子，描述形状信息的Hough变换, 傅里叶形状描述符, SIFT[27], SURF[28], Haar等特征。文献[8,16,20–27]均是提取图像的低级特征构建特征向量用于零样本学习中对象的视觉表达。

虽然基于传统计算机视觉的图像特征提取算法被学者们不断地创新改进，但是使用这些方法提取的特征仍然属于较低级的特征表达，无法与高层语义建立联系，存在较大的语义鸿沟。随着卷积神经网络和深度学习在计算机视觉领域取得巨大成果，如今对图像特征的提取更为有效的还是基于深度卷积神经网络的方法。深度卷积神经网络通过使用一系列的卷积核和非线性激活函数可以从原始图像中提取更高层级的抽象特征，而据学者研究表明，越抽象的描述越能表示图像的特征，这也是该方法明显优于传统方法的原因。文献[16]使用GRNN网络训练视觉属性到对象类别的分类器，大幅提升了预测精度。文献[29]分析基于视觉属性的DAP式零样本学习，设计了一个端到端的网络。自2014年以来基于深度卷积神经网络的方法被广泛应用于零样本

学习，至此在零样本学习中提取的视觉特征也开始由浅入深、语义相关性由低到高。

深度卷积神经网络自提出以来就不断飞速发展，优秀的网络也是层出不穷，目前在零样本学习中使用较为广泛主要有VGG, GoogleNet和ResNet，表2给出了近5年内出现在AAAI, NIPS和CVPR等顶级国际会议中关于零样本学习使用3种深度卷积神经网络的论文数量统计表。

**表 2 零样本学习中深度卷积神经网络使用情况统计表**

| 网络 | 论文数量 |
|---|---|
| VGG | 501 |
| GoogleNet | 271 |
| ResNet | 397 |

根据表2可以看出，目前VGG是在零样本学习中使用最为广泛，ResNet次之，使用GoogleNet的文章数量相对较少。这主要是因为由于计算机算力的限制，很多学者没有昂贵的专业深度学习计算资源，所以在网络选择上受到硬件成本的限制，所以相对网络结构较小，对计算能力需求较低的VGG-16和VGG-19的使用率相对较高。

GoogleNet比VGG具有更多的层数，提取的特征更加抽象，而且Inception模块使GoogleNet可以获取更多不同尺度的视觉信息，且该模块至今仍在不断更新；ResNet通过残差学习，构建更深层次的网络，可以达到比VGG和GoogleNet更好的预测精度。所以，在硬件成本允许的情况下，选择使用更深层的网络将会对最终模型的识别精度起到更好的提升作用。

### 3.2 语义特征

零样本学习之所以可以完成传统监督学习无法完成的对未知类识别的任务，关键因素就在于零样本学习除了将视觉特征用于识别外，还引入了语义特征，从而超越了互斥对象类之间的类边界。在零样本学习中，语义特征的提取一般独立于视觉特征的提取，不过近年来也有许多学者提出了端到端的网络[30–32]统一了这两个过程。

语义特征提取的目的在于构建语义向量，它必须同时具有足够的语义性和较强的判断性。两者相互矛盾，语义性强要求语义向量能够包含更多对象的细节信息，其中包含了与分类任务不相关的信息，如动物的习性、人类的习惯、场景的描述等；而判断性强则要求语义向量更应该关注最能区分不同对象类别的信息。如何平衡两者之间的权重，强化零样本学习的泛化能力也是目前零样本学习领域的研究重点之一。

零样本学习中主要的语义特征提取方法分为基于属性、基于词嵌入向量和基于知识图。

基于属性的方法是根据专家知识人工标注的类标签语义向量，它是一种对象之间可共享的自然语言属性。它是最常用也是最有效的语义特征构建方式。目前对象类属性的获取方式有3种，(1)认知科学已有的研究成果，如AwA使用了认知科学中已有的85个高质量语义属性描述不同动物类；(2)专业领域的专家知识，如CUB[33]使用行业专家定义的312个细粒度的鸟类特征属性构造而来；(3) AMT对属性进行众包学习，在众包网站AMT上获取由广大群众共同决定的属性。

一般在获取完属性后，还需要对每个对象类的各个属性进行人工打分，进一步估计对象类与属性间的关联强度，图4展示了AwA数据集中部分类与属性的关联度，颜色越深，关联度越小。

基于词嵌入向量的方法是使用自然语言处理技术自动地处理大规模、更易获得的无监督文本文档库，如Wikipedia, Google News等，从而构造语义向量。它的表达形式有one-hot编码, CBOW, Skip-gram和GloVe。其中one-hot编码是离散式词向量，它形式简单，具有鲁棒性，但也存在着易引发维度灾难和无法避免语义鸿沟的缺陷；CBOW, Skip-gram和GloVe均属于分布式表达[34]，是连续式词向量，它维度较小，且语义相似的词距离相近，是目前最成功的词向量表达形式。

基于知识图的方法[32,35]不同于以上两种通过语义向量间接建立对象类间关联关系，知识图是直接使用类间相似度构建语义向量，向量的每个元素表示该类与其它类的相似度，在零样本学习中使用大规模词汇词典WordNet为对象类构建层次化相似度。WordNet是一个覆盖范围宽广的英语词汇语义网，名词、动词、形容词和副词各自被组织成一个同义词的网络，每个同义词集合都代表一个基本的语义概念，并且这些集合之间也由各种关系连接，是一个具有层次化结构的网络。因此，对于所有未知类与已知类，可以抽取他们在WordNet中最小公共上位子集，使用路径长度计算每个类与上位子集中所有类的相似度，利用这些相似度构建具有层次化特点的语义向量。

在零样本学习模型已经发展到相对稳定，学者开始细粒度研究模型中各个模块的创新，目前视觉特征提取已经普遍采用基于深度卷积神经网络的方式获取，而语义特征的提取除了以上3种常用方法，学者们仍在不断提出新的方法：(1)出于类的视觉信息与图像本身的视觉特征结构更一致的考虑，学者们提出了可以在零样本学习中使用对象类

的各种不完美的、带噪声的图像数据[36]或者是人类目光注视数据[37]，这一类由原始图像构成的视觉语义向量；(2)为了充分利用文本信息提高细粒度对象类识别的性能，文献[38]使用对图像更加精细化的句子描述和文献[39]使用类的原始文本文档信息，构建基于文本文档的语义词向量。虽然学者们不断的提出更多的创新方法，但是目前关注度最高的还是分布式词向量，它在大规模物体识别上具有较强的灵活性和可扩展性。

### 3.3 视觉-语义映射

视觉-语义映射是解决零样本学习问题必不可少的基石，是图像特征与语义向量之间的连接的枢纽。一旦建立好视觉-语义映射，便可以计算任意未知类数据和未知类原型之间的相似度，并基于该相似度对未知类进行分类。

有3种方式构建视觉-语义映射：(1)正向映射，在视觉端构建，将图片特征映射到语义向量空间，并在语义向量空间进行未知类识别；(2)反向映射，在语义端构建，将语义向量映射到图片特征空间，并在图片特征空间进行未知类识别；(3)公共映射，在公共端构建，将图片特征和语义向量映射到公共空间，并在公共空间内完成识别任务，3种映射方式如图5所示，其中$f$表示映射函数，$f_s$表示语义空间映射函数，$f_v$表示视觉空间映射函数。

正向映射从2009年开始一直沿用至今，是零样本学习的主流映射方式，如Baroni等人[40]使用线性脊回归模型实现线性正向映射，Wang等人[41]使用贝叶斯网络实现非线性正向映射；公共映射是在2013年以双线性方程的形式首次被提出[42]，该映射方式具有良好的判决性和泛化能力，也已被学者们

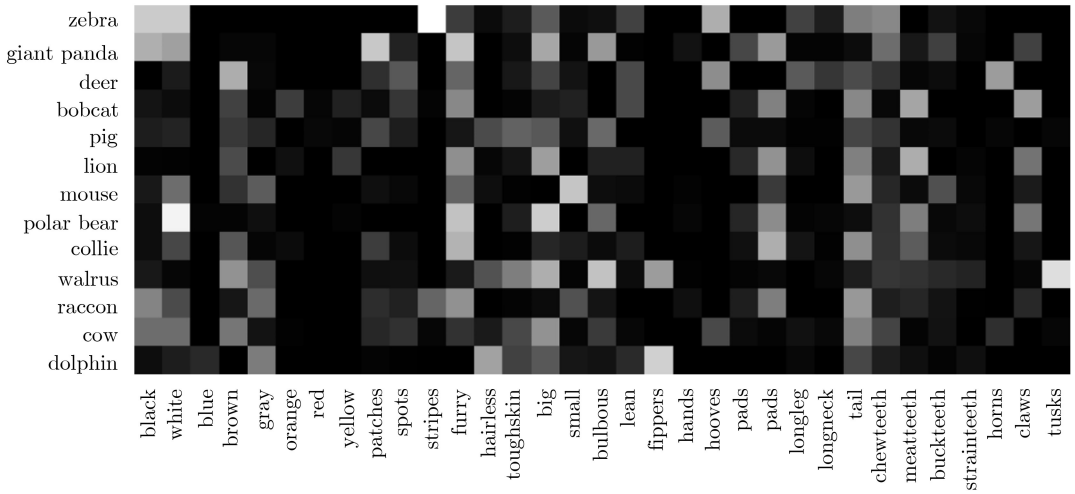



图 4 AwA类-属性关系矩阵[7]

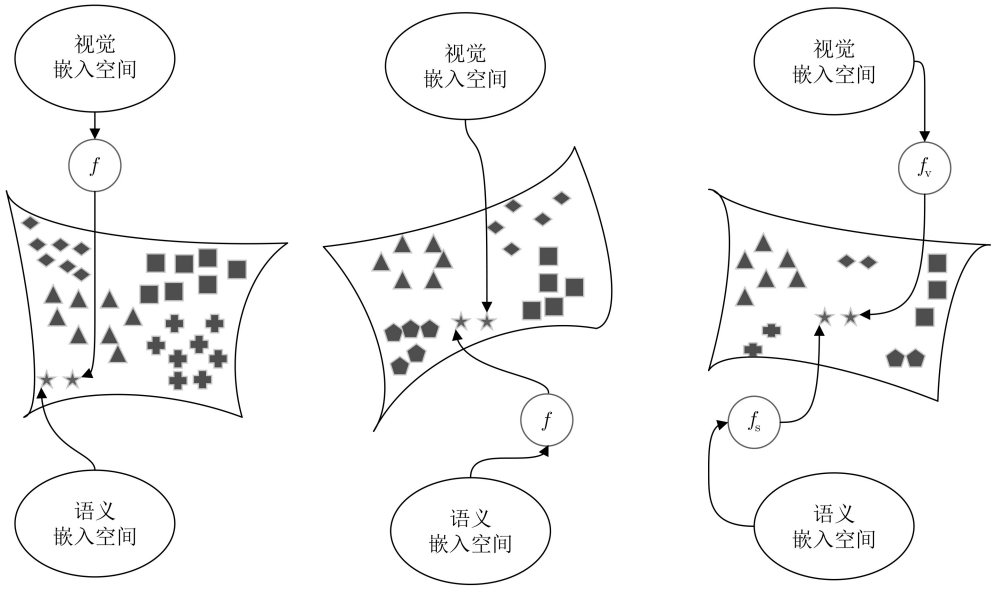



图 5 3种视觉-语义映射示意图

广泛采纳，如SJE模型、DA-SJE模型、MultiCUE模型等，此外，Romera-Paredes等人[29]在双线性基础上使用平方损失函数和参数规整化调整，让模型的执行更简单高效，目前已成为零样本学习中基准算法之一；反向映射从2015年之后才被提出，它可以用来为未知类构建虚拟的训练样本，从而允许在图片特征空间内直接进行有监督分类学习，如文献[43]中使用DAE, GAN等模型将未知类属性向量映射到图像特征空间，利用SVM进行有监督分类。

由于基于深度学习技术提取的图片特征具有良好的结构性，同类数据紧凑分布，判决能力较强，所以反向映射的使用率正逐年上升，在3种映射方式中发展空间最大。

## 4 性能评估

### 4.1 数据集

在零样本学习领域有7个常用数据集，它们分别是AwA, CUB, aPY, SUN, Flower, Dogs和ImageNet。其中前6个数据集为小规模数据集，最后一个为大规模数据集。

AwA (Animal with Attribute)[7]是动物图片数据集，总共50类，40类为训练集，10类为测试集，类别语义表示85维，共30475张图片。但是由于版本权限问题，AwA中的图片无法获取，所以作者便推出了AwA2[13]版本数据集，并将图片数量增加到37322张。

CUB (Caltech-UCSD-Birds-200-2011)[33]是来自于专业鸟类网站数据构成的鸟类图片数据集，属于细粒度分类数据集，总共200类，150类为训练集，50类为测试集，类别语义表示为312维，共11788张图片。

aPY (aPascal-aYahoo)[44]包含2个子集，分别是由PASCAL VOC 2008数据集中的20个物体类子集和来自Yahoo图片搜索引擎的12个物体类组成，总共32个类别，aPascal中的20类为训练集，aYahoo中的12类为测试集，类别语义表示为64维，共15339张图片。

SUN (SUN attribute dataset)[45,46]涵盖各种环境场景和场景内物体的注释图像数据集，主要用于细粒度的场景分类问题，总共717个类别，707类为训练集，10类为测试集，每个类别20张图片，类别语义表示为102维。

Flower (Oxford Flower Dataset)[47]由牛津大学创建的英国花卉数据集，适用于细粒度分类问题。该数据集提供两个版本，一种是包含17种花卉的Flower17[48]和包含102种花卉的Flower102[49]。由于Flower17数据量太少，在深度卷积网络中容易出现过拟合现象，所以零样本学习中一般选用Flower102作为数据集，其中随机选取82个类用于训练，20个类用于测试。

Dogs (Stanford Dogs Dataset)[50]包含了世界各地120种犬类的图片，属于细粒度分类数据集，总共120类，100类为训练集，20类为测试集，共20580张图片。

ImageNet[51]是深度学习中使用最广泛的大规模数据集，有超过1400万的标注图像，是世界上图像识别最大数据库。由于完整的ImageNet太过庞大，一般应用于零样本学习中的数据集只是部分ImageNet，如ImageNet2010 1K, ImageNet 2011 21K和ImageNet2012 1K。

### 4.2 指标与方案

#### 4.2.1 评价指标

根据使用的数据集规模不同，零样本学习性能的评价指标也不同。

针对小规模数据集而言，主要的评价指标是分类精确度(acc)、精确率(precision)和召回率(recall)，三者分别从不同角度对零样本学习性能进行评价。以分类任务为例，给出四类分类结果的符号表示，正确预测正样本(True Positive, TP)、错误预测正样本(False Positive, FP)、正确预测负样本(True Negative, TN)、错误预测负样本(False Negative, FN)，则

分类精确率：分类器正确分类的样本数与总样本数之比，数值越高，性能越好，

$$\text{acc}=\frac{\text{TN}+\text{TP}}{\text{TP}+\text{TN}+\text{FN}+\text{FP}} \tag{2}$$

精确率：预测为正样本的样本中真正为正样本所占的比例，

$$\text{precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{3}$$

召回率：正样本中被成功预测为正样本的比例，

$$\text{recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{4}$$

此外，还可以使用AUC来综合评估算法性能，通过面积直观地表示分类性能的好坏；或者当精确率与召回率发生矛盾时，也可以使用F-Measuer对precision和recall加权调和平均，如式(5)所示，若$\alpha=1$，则是常见的F1指标。

$$F=\frac{(\alpha^2+1)\text{Precision}\cdot\text{Recall}}{\alpha^2(\text{Precision}+\text{Recall})} \tag{5}$$

针对大规模数据集而言，使用的评价指标为Flat hit @ k(f@k)和Hierarchical precision @ k(hp@k)。

f@k：模型预测的前$k$个类标签中含有真实类

标签的测试数据在整个测试集中的百分比。特别地，f@1等价于传统的分类精确率。

Hp@k：模型预测的前$k$个类标签中与真实类标签的$k$个最近邻中的平均重叠度。具体来说，对于任意真实标签，首先从ImageNet的标签层级中选取$k$个最近邻，其次计算它与模型预测的$k$个标签的重叠程度。

#### 4.2.2 传统零样本学习VS泛化零样本学习

传统零样本学习(conventional Zero-Shot Learning, cZSL)是指在实验阶段仅使用已知类训练模型，仅使用未知类测试模型；泛化零样本学习(generalized Zero-Shot Learning, gZSL)则不再将测试数据强制认定为仅来自未知类，而是对测试数据的来源做更松弛化的假设(测试数据可以来自于所有类别中的任意对象类)，两者是目前两种主流的实验评价配置方案。

一个切实有用的零样本学习系统应该既能精确地识别已知类，又能准确地区分未知类。它不但可以区分待识别对象是来自于已知类还是未知类，也可以识别待识别对象属于已知类或未知类中的哪种物体。但是由于传统零样本学习的测试阶段设置过于严苛，不能真实反映现实世界中物体识别的情景，为了更好地评价零样本学习模型在实际情况中的性能，泛化零样本学习必不可少[52]。

近年来，关于泛化零样本学习的研究逐渐被学者重视，2016年由Chao等人[52]首次提出，Song等人[53]提出首个用在广义设置问题的直推式模型QFSL (Quasi-Fully Supervised Learning)，有效解决了零样本学习中的分类偏差问题。

### 4.3 主流模型性能对比分析

本文选取几个典型的和最新的零样本学习模型分别在传统零样本学习和泛化零样本学习设置下进行模型性能对比，为了便于统计各个模型的性能数据，本文选取了在零样本学习中使用最为广泛的3个数据集，分别为AwA, CUB和SUN。对比数据来源于引用算法的原始数据或其他论文实现并公布的结果，具体如表3所示。

传统零样本学习性能比较选取两个量化指标对模型性能进行分析，其中SS表示数据集在标准分割规则下得到的识别精度，PS表示数据集在文献[13]的分割规则下得到的识别精度。从对比结果可以发现：(1)早期基于浅层图像特征的模型性能普遍较低，当使用深度卷积神经网络提取图像的深度特征后，性能大幅度提升；(2)零样本学习对于粗粒度数据集的性能要优于细粒度数据集，在表中表现为各个模型在AwA数据集上的性能均大于在CUB和SUN数据集上的性能。

泛化零样本学习性能比较选取3个量化指标对模型性能进行分析，其中$U \to T$表示假设测试数据来自于未知类，但在所有对象类中进行识别，$S \to T$表示假设测试数据来自于已知类，但在所有对象类中进行识别，$H$表示调和平均数。从对比结果可以发现：(1)大多数模型在泛化零样本学习设置下的性能指标都不理想，这说明了传统的零样本学习设置存在一定程度的缺陷，未能较真实地契合现实世界的分类情况；(2)QFSL模型在泛化能力上要明显好于其他模型，由于该模型属于直推式零样本学习模型，所以从侧面说明了只依靠已知类数据构建的算法缺乏泛化能力。

结合表3两组数据不难发现，QFSL在传统零样本学习和泛化零样本学习设置中均有较好的表现，而且文献[54]通过对比多种模型证明了采用直推式

**表 3　零样本学习性能比较(%)**

| 方法 | 传统零样本学习 | | | | | | 泛化零样本学习 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AwA | | CUB | | SUN | | AwA | | | CUB | | | SUN | | |
| | SS | PS | SS | PS | SS | PS | $U\to T$ | $S\to T$ | $H$ | $U\to T$ | $S\to T$ | $H$ | $U\to T$ | $S\to T$ | $H$ |
| IAP | 46.9 | 35.9 | 27.1 | 24.0 | 17.4 | 19.4 | 0.9 | 87.6 | 1.8 | 0.2 | **72.8** | 0.4 | 1.0 | 37.8 | 1.8 |
| DAP | 58.7 | 46.1 | 37.5 | 40.0 | 38.9 | 39.9 | 0.0 | 84.7 | 0.0 | 1.7 | 67.9 | 3.3 | 4.2 | 25.1 | 7.2 |
| DeViSE | 68.6 | 59.7 | 53.2 | 52.0 | 57.5 | **56.5** | **17.1** | 74.7 | **27.8** | **23.8** | 53.0 | 32.8 | **16.9** | 27.4 | **20.9** |
| ConSE | 67.9 | 44.5 | 36.7 | 34.3 | 44.2 | 38.8 | 0.5 | **90.6** | 1.0 | 1.6 | 72.2 | 3.1 | 6.8 | 39.9 | 11.6 |
| SJE | 69.5 | **61.9** | **55.3** | 53.9 | 57.1 | 53.7 | 8.0 | 73.9 | 14.4 | 23.5 | 59.2 | **33.6** | 14.7 | 30.5 | 19.8 |
| SAE | **80.7** | 54.1 | 33.4 | 33.3 | 42.4 | 40.3 | 1.1 | 82.2 | 2.2 | 7.8 | 54.0 | 13.6 | 8.8 | 18.0 | 11.8 |
| SYNC | 71.2 | 46.6 | 54.1 | **55.6** | **59.1** | 56.3 | 10.0 | 90.5 | 18.0 | 11.5 | 70.9 | 19.8 | 7.9 | **43.3** | 13.4 |
| LDF | 83.4 | – | **70.4** | – | – | – | – | – | – | – | – | – | – | – | – |
| SP-AEN | – | 58.5 | – | 55.4 | – | **59.2** | 23.3 | 90.9 | 37.1 | 34.7 | 70.6 | 46.6 | 24.9 | **38.6** | 30.3 |
| QFSL | **84.8** | **79.7** | 69.7 | 72.1 | **61.7** | 58.3 | **66.2** | **93.1** | **77.4** | **71.5** | **74.9** | **73.2** | **51.3** | 31.2 | **38.8** |

方法通常可以获得更高的性能。虽然现阶段使用归纳式方法的模型较多，但是随着零样本学习对泛化能力的需求不断增加，未来基于直推式的模型将会越来越多的被学者们提出。

## 5 技术难题

零样本学习是在传统监督学习上发展而来，但又不同于传统监督学习，它主要是利用已知类训练模型，借助知识迁移实现对未知类的识别。所以除了存在传统监督学习中固有的过拟合问题外，学者们在研究过程中发现零样本学习所特有的新问题，主要有3种。

(1) 领域漂移：针对基于属性构建语义向量的模型而言，当同一种属性在不同的类别中，视觉特征表达可能差别很大[55]。如图6所示的斑马和猪都有尾巴的属性，但是两者的尾巴的视觉特征相差很远，如果用斑马训练出来的模型将很难正确识别猪。目前最流行的解决办法是将映射到语义空间的样本再重建回去，这样学习到的映射就能保留更多的信息。如文献[56]提出SAE模型，通过在岭回归基础上构建一个简单的自编码器，从而解决领域漂移问题。

(2) 枢纽点：在高维空间中，某些点会成为大多数点的最近邻点，这是高维空间中固有的问题。由于一般的零样本学习模型最终分类使用的是K-NN算法，所以枢纽点问题严重影响分类结果。该问题的解决方案有两个，一是若模型采用岭回归方式建立，则需要将视觉-语义映射方式改为反向映射，文献[57]证明了基于岭回归的方法会加重该问题，采用改变映射方式可以帮助解决枢纽点问题；二是使用生成模型(自编码器[56,58]和GAN等)生成测试集的样本，将零样本学习转变成传统的监督分类问题，避免K-NN操作，从而消除枢纽点问题。

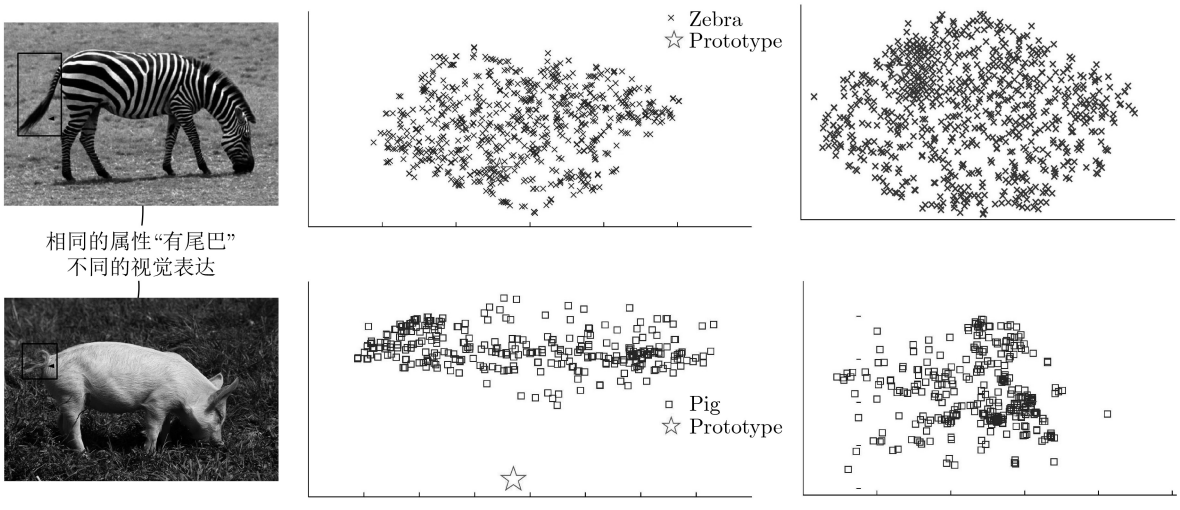
(3) 语义间隔：由于表示视觉特征的图像与表示非视觉特征的语义经过特征提取之后，图像特征空间与语义空间中对应类别的流形不一致，导致直接学习两者之间的映射较为困难，从而引发语义间隔问题。如图7所示，Li等人[59]提出由于$X_s$和$K_s$的流形不同，很难实现$X_s \to K_s$的映射，从而造成语义间隔，所以通过将$K_s$转换为与$X_s$具有相同流形的$\tilde{K}_s$再做$X_s \to \tilde{K}_s$的映射。

除了以上3种目前研究的热点问题，零样本学习还存在一些诸如语义丢失问题[60]、样本通过映射坍塌到一点的问题[14]等。针对零样本学习中存在的问题，一般可以考虑尝试不同的视觉-语义映射方式、实现端到端的模型结构，他们可以帮助避免一些问题的发生。虽然目前语义向量提取的来源已经固定，属性比词向量的性能好，词向量比属性易获得，但是如果可以设计出兼顾两者优点的语义向量空间，那么这对解决零样本学习的技术难题将提供极大的帮助。

## 6 结束语

本文主要从基本介绍、关键技术、性能评估和技术难题4个方面详细地介绍了零样本学习，并结合现有文献分析了关键技术的未来发展趋势，列举了现有零样本学习中的常用数据集及其特点，对比了现有零样本学习中几类具有代表性模型的性能，总结了目前零样本学习面临的主要难题和相应的解决办法。

理论上，凡是面临着样本数据缺失问题的领域，均适合用零样本学习来解决。例如，将零样本




(a) 视觉空间 (b) 属性空间 (c) 嵌入空间

图 6 领域漂移示例图[55]

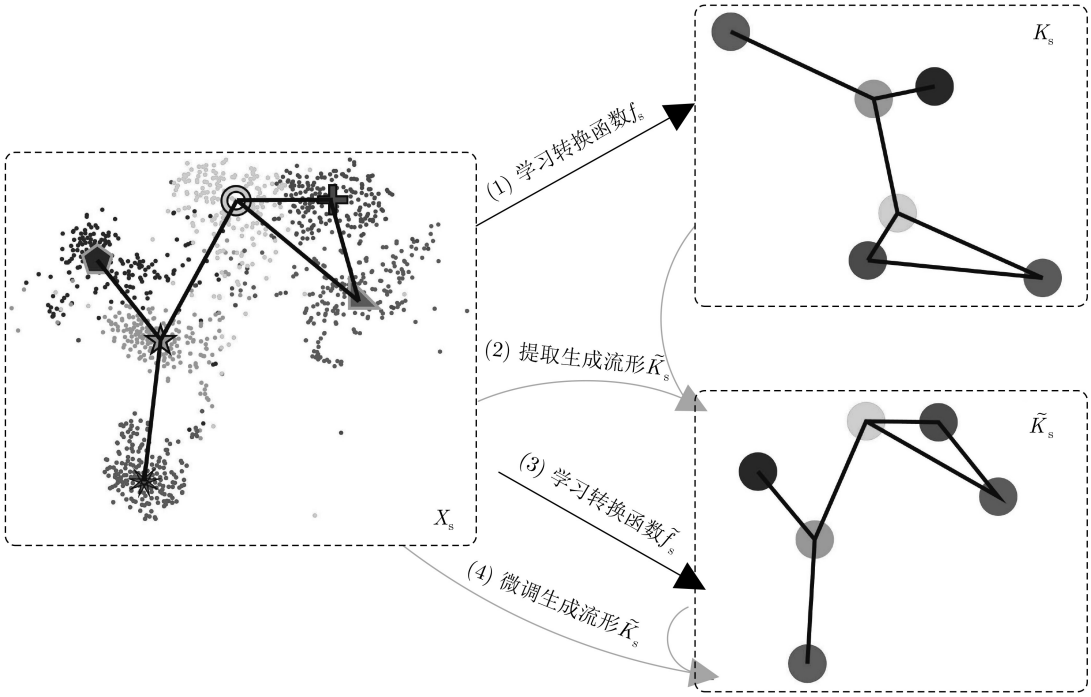



图 7 语义间隔示例图

学习运用在AwA, CUB数据集上，做到对珍稀动物的识别为人们带来生态效益；在SUN数据集上，训练模型识别从未看见过的场景，提升自动无人驾驶对周围环境的认知能力；将其运用在垃圾分类上，也将会实现更加智能化，更加精确的垃圾分类模型。

零样本学习领域拥有着巨大的潜力，同时伴随着很多挑战等待学者们的进一步研究，通过对现有文献的思考，本文列出以下未来零样本学习的研究方向：(1)在泛化零样本学习设置下如何提高模型的精确度和泛化能力；(2)如何提高细粒度零样本学习的识别精度；(3)怎样构建端到端的零样本学习模型；(4)如何联合少样本学习、开集识别等相关课题完善零样本学习；(5)如何逐步实现终身学习的目标。

## 参 考 文 献


[1] SUN Yi, CHEN Yuheng, WANG Xiaogang, *et al*. Deep learning face representation by joint identification-verification[C]. The 27th International Conference on Neural Information Processing Systems, Montreal, Canada, 2014: 1988–1996.

[2] LIU Chenxi, ZOPH B, NEUMANN M, *et al*. Progressive neural architecture search[C]. The 15th European Conference on Computer Vision, Munich, Germany, 2018: 19–35.

[3] LEDIG C, THEIS L, HUSZÁR F, *et al*. Photo-realistic single image super-resolution using a generative adversarial network[C]. The IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 105–114.

[4] BIEDERMAN I. Recognition-by-components: A theory of human image understanding[J]. *Psychological Review*, 1987, 94(2): 115–147. doi: 10.1037/0033-295X.94.2.115.

[5] LAROCHELLE H, ERHAN D, and BENGIO Y. Zero-data learning of new tasks[C]. The 23rd National Conference on Artificial Intelligence, Chicago, USA, 2008: 646–651.

[6] PALATUCCI M, POMERLEAU D, HINTON G, *et al*. Zero-shot learning with semantic output codes[C]. The 22nd International Conference on Neural Information Processing Systems, Vancouver, Canada, 2009: 1410–1418.

[7] LAMPERT C H, NICKISCH H, and HARMELING S. Learning to detect unseen object classes by between-class attribute transfer[C]. The IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009: 951–958. doi: 10.1109/CVPR.2009.5206594.

[8] HARRINGTON P. Machine Learning in Action[M]. Greenwich, CT, USA: Manning Publications Co, 2012: 5–14.

[9] ZHOU Dengyong, BOUSQUET O, LAL T N, *et al*. Learning with local and global consistency[C]. The 16th International Conference on Neural Information Processing Systems, Whistler, Canada, 2003: 321–328.

[10] 刘建伟, 刘媛, 罗雄麟. 半监督学习方法[J]. 计算机学报, 2015, 38(8): 1592–1617. doi: 10.11897/SP.J.1016.2015.01592.
LIU Jianwei, LIU Yuan, and LUO Xionglin. Semi-supervised learning methods[J]. *Chinese Journal of Computers*, 2015, 38(8): 1592–1617. doi: 10.11897/

SP.J.1016.2015.01592.

[11] SUNG F, YANG Yongxin, LI Zhang, *et al*. Learning to compare: Relation network for few-shot learning[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 1199–1208.

[12] FU Yanwei, XIANG Tao, JIANG Yugang, *et al*. Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content[J]. *IEEE Signal Processing Magazine*, 2018, 35(1): 112–125. doi: 10.1109/MSP.2017.2763441.

[13] XIAN Yongqin, LAMPERT C H, SCHIELE B, *et al*. Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 41(9): 2251–2265. doi: 10.1109/TPAMI.2018.2857768.

[14] WANG Wenlin, PU Yunchen, VERMA V K, *et al*. Zero-shot learning via class-conditioned deep generative models[C]. The 32nd AAAI Conference on Artificial Intelligence, New Orleans, USA, 2018: 4211–4218.

[15] FU Yanwei, HOSPEDALES T M, XIANG Tao, *et al*. Attribute learning for understanding unstructured social activity[C]. The 12th European Conference on Computer Vision, Florence, Italy, 2012: 530–543.

[16] ANTOL S, ZITNICK C L, and PARIKH D. Zero-shot learning via visual abstraction[C]. The 13th European Conference on Computer Vision, Zurich, Switzerland, 2014: 401–416.

[17] ROBYNS P, MARIN E, LAMOTTE W, *et al*. Physical-layer fingerprinting of LoRa devices using supervised and zero-shot learning[C]. The 10th ACM Conference on Security and Privacy in Wireless and Mobile Networks, Boston, USA, 2017: 58–63. doi: 10.1145/3098243.3098267.

[18] YANG Yang, LUO Yadan, CHEN Weilun, *et al*. Zero-shot hashing via transferring supervised knowledge[C]. The 24th ACM international conference on Multimedia, Amsterdam, The Netherlands, 2016: 1286–1295. doi: 10.1145/2964284.2964319.

[19] PACHORI S, DESHPANDE A, and RAMAN S. Hashing in the zero shot framework with domain adaptation[J]. *Neurocomputing*, 2018, 275: 2137–2149. doi: 10.1016/j.neucom.2017.10.061.

[20] LIU Jingen, KUIPERS B, and SAVARESE S. Recognizing human actions by attributes[C]. The IEEE Conference on Computer Vision and Pattern Recognition, Colorado, USA, 2011: 3337–3344.

[21] FU Yanwei, HOSPEDALES T M, XIANG Tao, *et al*. Learning multimodal latent attributes[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 36(2): 303–316. doi: 10.1109/TPAMI.2013.128.

[22] JAIN M, VAN GEMERT J C, MENSINK T, *et al*. Objects2action: Classifying and localizing actions without any video example[C]. The IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 4588–4596.

[23] XU Baohan, FU Yanwei, JIANG Yugang, *et al*. Video emotion recognition with transferred deep feature encodings[C]. The 2016 ACM on International Conference on Multimedia Retrieval, New York, USA, 2016: 15–22.

[24] JOHNSON M, SCHUSTER M, LE Q V, *et al*. Google's multilingual neural machine translation system: Enabling zero-shot translation[J]. *Transactions of the Association for Computational Linguistics*, 2017, 5: 339–351. doi: 10.1162/tacl_a_00065.

[25] PRATEEK VEERANNA S, JINSEOK N, ENELDO L M, *et al*. Using semantic similarity for multi-label zero-shot classification of text documents[C]. The 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, 2016: 423–428.

[26] DALAL N and TRIGGS B. Histograms of oriented gradients for human detection[C]. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, USA, 2005: 886–893.

[27] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. *International Journal of Computer Vision*, 2004, 60(2): 91–110. doi: 10.1023/B:VISI.0000029664.99615.94.

[28] BAY H, ESS A, TUYTELAARS T, *et al*. Speeded-up robust features (SURF)[J]. *Computer Vision and Image Understanding*, 2008, 110(3): 346–359. doi: 10.1016/j.cviu.2007.09.014.

[29] ROMERA-PAREDES B and TORR P H S. An embarrassingly simple approach to zero-shot learning[C]. The 32nd International Conference on International Conference on Machine Learning, Lille, France, 2015: 2152–2161.

[30] ZHANG Li, XIANG Tao, and GONG Shaogang. Learning a deep embedding model for zero-shot learning[C]. The IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 3010–3019.

[31] LI Yan, ZHANG Junge, ZHANG Jianguo, *et al*. Discriminative learning of latent features for zero-shot recognition[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 7463–7471.

[32] WANG Xiaolong, YE Yufei, and GUPTA A. Zero-shot recognition via semantic embeddings and knowledge graphs[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018:

6857–6866.

[33] WAH C, BRANSON S, WELINDER P, *et al*. The caltech-UCSD birds-200-2011 dataset[R]. Technical Report CNS-TR-2010-001, 2011.

[34] MIKOLOV T, SUTSKEVER I, CHEN Kai, *et al*. Distributed representations of words and phrases and their compositionality[C]. The 26th International Conference on Neural Information Processing Systems, Lake Tahoe, USA, 2013: 3111–3119.

[35] LEE C, FANG Wei, YEH C K, *et al*. Multi-label zero-shot learning with structured knowledge graphs[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 1576–1585.

[36] JETLEY S, ROMERA-PAREDES B, JAYASUMANA S, *et al*. Prototypical priors: From improving classification to zero-shot learning[J]. arXiv: 2015, 1512.01192.

[37] KARESSLI N, AKATA Z, SCHIELE B, *et al*. Gaze embeddings for zero-shot image classification[C]. The IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 6412–6421.

[38] REED S, AKATA Z, LEE H, *et al*. Learning deep representations of fine-grained visual descriptions[C]. The IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 49–58.

[39] ELHOSEINY M, ZHU Yizhe, ZHANG Han, *et al*. Link the head to the "beak": Zero shot learning from noisy text description at part precision[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 6288–6297. doi: 10.1109/CVPR.2017.666.

[40] LAZARIDOU A, DINU G, and BARONI M. Hubness and pollution: Delving into cross-space mapping for zero-shot learning[C]. The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 2015: 270–280.

[41] WANG Xiaoyang and JI Qiang. A unified probabilistic approach modeling relationships between attributes and objects[C]. The IEEE International Conference on Computer Vision, Sydney, Australia, 2013: 2120–2127.

[42] AKATA Z, PERRONNIN F, HARCHAOUI Z, *et al*. Label-embedding for attribute-based classification[C]. The IEEE Conference on Computer Vision and Pattern Recognition, Portland, USA, 2013: 819–826.

[43] JURIE F, BUCHER M, and HERBIN S. Generating visual representations for zero-shot classification[C]. The IEEE International Conference on Computer Vision Workshops, Venice, Italy, 2017: 2666–2673.

[44] FARHADI A, ENDRES I, HOIEM D, *et al*. Describing objects by their attributes[C]. 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009: 1778–1785. doi: 10.1109/CVPR.2009.5206772.

[45] PATTERSON G, XU Chen, SU Hang, *et al*. The sun attribute database: Beyond categories for deeper scene understanding[J]. *International Journal of Computer Vision*, 2014, 108(1/2): 59–81.

[46] XIAO Jianxiong, HAYS J, EHINGER K A, *et al*. Sun database: Large-scale scene recognition from abbey to zoo[C]. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, USA, 2010: 3485–3492. doi: 10.1109/CVPR.2010.5539970.

[47] NILSBACK M E and ZISSERMAN A. Delving deeper into the whorl of flower segmentation[J]. *Image and Vision Computing*, 2010, 28(6): 1049–1062. doi: 10.1016/j.imavis.2009.10.001.

[48] NILSBACK M E and ZISSERMAN A. A visual vocabulary for flower classification[C]. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, USA, 2006: 1447–1454. doi: 10.1109/CVPR.2006.42.

[49] NILSBACK M E and ZISSERMAN A. Automated flower classification over a large number of classes[C]. The 6th Indian Conference on Computer Vision, Graphics & Image Processing, Bhubaneswar, India, 2008: 722–729. doi: 10.1109/ICVGIP.2008.47.

[50] KHOSLA A, JAYADEVAPRAKASH N, YAO Bangpeng, *et al*. Novel dataset for fine-grained image categorization: Stanford dogs[C]. CVPR Workshop on Fine-Grained Visual Categorization, 2011.

[51] DENG Jia, DONG Wei, SOCHER R, *et al*. ImageNet: A large-scale hierarchical image database[C]. 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009: 248–255.

[52] CHAO Weilun, CHANGPINYO S, GONG Boqing, *et al*. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild[C]. The 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 2016: 52–68.

[53] SONG Jie, SHEN Chengchao, YANG Yezhou, *et al*. Transductive unbiased embedding for zero-shot learning[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 1024–1033.

[54] 李亚南. 零样本学习关键技术研究[D]. [博士论文], 浙江大学, 2018: 40–43.
LI Yanan. Research on key technologies for zero-shot learning[D]. [Ph.D. dissertation], Zhejiang University, 2018: 40–43

[55] FU Yanwei, HOSPEDALES T M, XIANG Tao, *et al*. Transductive multi-view zero-shot learning[J]. *IEEE*

*Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(11): 2332–2345. doi: 10.1109/TPAMI.2015.2408354.

[56] KODIROV E, XIANG Tao, and GONG Shaogang. Semantic autoencoder for zero-shot learning[C]. The IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 4447–4456.

[57] STOCK M, PAHIKKALA T, AIROLA A, *et al*. A comparative study of pairwise learning methods based on kernel ridge regression[J]. *Neural Computation*, 2018, 30(8): 2245–2283. doi: 10.1162/neco_a_01096.

[58] ANNADANI Y and BISWAS S. Preserving semantic relations for zero-shot learning[J]. arXiv: 2018, 1803.03049.

[59] LI Yanan, WANG Donghui, HU Huanhang, *et al*. Zero-shot recognition using dual visual-semantic mapping paths[C]. The IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 5207–5215.

[60] CHEN Long, ZHANG Hanwang, XIAO Jun, *et al*. Zero-shot visual recognition using semantics-preserving adversarial embedding networks[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 1043–1052.

兰 红：女，1969年生，教授，硕士生导师，主要研究方向为计算机视觉、图像处理与模式识别.

方治屿：男，1993年生，硕士生，研究方向为计算机视觉与深度学习.